VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks

Year: 2023 | Citations: 608 | Authors: Wen Wang, Zhe Chen, Xiaokang Chen

---

Abstract

Large language models (LLMs) have notably accelerated progress towards artificial general intelligence (AGI), with their impressive zero-shot capacity for user-tailored tasks, endowing them with immense potential across a range of applications. However, in the field of computer vision, despite the availability of numerous powerful vision foundation models (VFMs), they are still restricted to tasks in a pre-defined form, struggling to match the open-ended task capabilities of LLMs. In this work, we present an LLM-based framework for vision-centric tasks, termed VisionLLM. This framework provides a unified perspective for vision and language tasks by treating images as a foreign language and aligning vision-centric tasks with language tasks that can be flexibly defined and managed using language instructions. An LLM-based decoder can then make appropriate predictions based on these instructions for open-ended tasks. Extensive experiments show that the proposed VisionLLM can achieve different levels of task customization through language instructions, from fine-grained object-level to coarse-grained task-level customization, all with good results. It's noteworthy that, with a generalist LLM-based framework, our model can achieve over 60\% mAP on COCO, on par with detection-specific models. We hope this model can set a new baseline for generalist vision and language models. The demo shall be released based on https://github.com/OpenGVLab/InternGPT. The code shall be released at https://github.com/OpenGVLab/VisionLLM.